How AI Systems Challenge the Conditions of Moral Agency?

Kalliokoski, Taina

Springer International Publishing AG
2020

Kalliokoski , T & Hallamaa , J 2020 , How AI Systems Challenge the Conditions of Moral Agency? in M Rauterberg (ed.) , Culture and Computing : 8th International Conference, C &C 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, þÿ D e n m a r k , J u l y 1 9 2 4 , 2 0 2 0 , P r o c e e d i n g s . L e c t u r e N o t e s i n C o m p u t e r S 12215 , Springer International Publishing AG , Cham , pp. 54-64 , International Conference, C &C 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, þÿ C o p e n h a g e n , D e n m a r k , J u l y 1 9 2 4 , 2 0 2 0 , C o p e n h a g e n , D e n m a r k , 1 9 /

http://hdl.handle.net/10138/332356
https://doi.org/10.1007/978-3-030-50267-6_5

unspecified
acceptedVersion

How AI Systems Challenge the Conditions of Moral Agency?[1]

Jaana Hallamaa[1][0000-0002-3897-4543] and Taina Kalliokoski[2][0000-0001-5683-6817]

[1] University of Helsinki, 00014 Helsinki, FINLAND
[2] University of Helsinki, 00014 Helsinki, FINLAND
jaana.hallamaa@helsinki.fi, taina.kalliokoski@helsinki.fi

Abstract. The article explores the effects increasing automation has on our conceptions of human agency. We conceptualize the central features of human agency as ableness, intentionality, and rationality and define responsibility as a central feature of moral agency. We discuss suggestions in favor of holding AI systems moral agents for their functions but join those who refute this view. We consider the possibility of assigning moral agency to automated AI systems in settings of machine-human cooperation but come to the conclusion that AI systems are not genuine participants in joint action and cannot be held morally responsible. Philosophical issues notwithstanding, the functions of AI systems change human agency as they affect our goal setting and pursuing by influencing our conceptions of the attainable. Recommendation algorithms on news sites, social media platforms, and in search engines modify our possibilities to receive accurate and comprehensive information, hence influencing our decision making. Sophisticated AI systems replace human workforce even in such demanding fields as medical surgery, language translation, visual arts, and composing music. Being second to a machine in an increasing number of fields of expertise will affect how human beings regard their own abilities. We need a deeper understanding of how technological progress takes place and how it is intertwined with economic and political realities. Moral responsibility remains a human characteristic. It is our duty to develop AI to serve morally good ends and purposes. Protecting and strengthening the conditions of human agency in any AI environment is part of this task.

Keywords: Moral Agency, AI-systems, Human-computer Interaction.

1 Introduction

Human dignity and moral responsibility are the two moral features central to our view of humanity. Respect for human dignity also forms one of the cornerstones of moral action. In Immanuel Kant's words, human beings must never be treated as means only but always also as ends in themselves [10]. The basic assumption that human beings are morally responsible for their deeds makes them legally accountable. Within societies in which human rights form the core of the legal thinking, historical developments tend to be inspired by these fundamental values.

An ever-widening variety of artificial intelligence (AI) systems plays a central role in our professional, economic, and social activities. There are high hopes that, in the near future, autonomous and adaptive robots and AI systems will take over a diverse

[1] This research was supported by the project Ethical AI for the Governance of Society (ETAIROS, grant #327352) funded by the Academy of Finland.

set of demanding or dangerous tasks. Complex webs of human–machine cooperation emerge as AI systems continuously collect information from their users and apply it as material for their learning algorithms, resulting in changes in their functioning in relation to their tasks and users.

Acting in interaction with an AI-modified environment may change the basic conditions of human agency and, therefore, affect the conditions of maintaining the intrinsic value of human dignity. In the following study, we explore how we can conceptualize the changes in human agency that a life closely tied to and connected with different types of AI are likely to bring forth.

2 Technologies are Designed for Instrumental Uses

Artifacts and technologies are designed for instrumental uses; as such, we evaluate their goodness—that is, their suitability, usefulness, and efficiency—in terms of the functions they were designed to serve [31] [32]. If a tool is useless in its instrumental task, we may scrupulously replace it with something better. It is not morally blameworthy to treat artifacts and technological systems simply and solely as a means to an end; as such, should an artifact or technology be ill suited for its use, we need not feel guilty for abandoning them and replacing them with something better. We may lament the lost time and resources, but as consumer legislation indicates, the objects of blame are the designers or manufacturers of the artifacts rather than the artifacts themselves. If there is any harm involved in the use of an artifact, we seek human actors on which to place the blame.

Currently, we are experiencing another period of hype related to the development of AI and its applications. National governments and multinational organizations anticipate that AI will not only save the economy but also provide solutions to all sorts of problems, ranging from the climate change crisis to elderly care. These expectations seem to rely on assigning a large role to AI systems in an increasing number of fields of human activity. [6] [16]

Autonomous and adaptive robots and AI systems are likely to take over a growing variety of tasks that are hazardous to human actors or so complicated that their performance involves a high probability of failure. [15] Rescue robots, for example, can accomplish life-saving actions that are too risky for human rescue workers; similarly, surgical robots can carry out an increasing number of different types of operations that require precision or durability that surpass human capacity. Most complicated calculations, data collection, and data management operations are already run by AI systems and would be impossible to perform without them. Energy, lighting, and water supply, maintenance, and delivery are also already run by AI-regulated systems. [16] The development of machine learning will multiply the number of fields that are based on AI-run procedures.

While AI helps to effectuate a vast number of human activities, it also renders human communities and societies extremely dependent on AI systems and their smooth functioning. Unlike mechanical machines, AI systems are not straightforward in their running, and the cause of their possible failure may be difficult to detect and both time-

consuming and expensive to fix. The probability of responsibility gaps—that is, situations in which it becomes impossible to assign moral responsibility for perilous outcomes on any specific agent—also increases with the broader use of AI systems [7]. The view of an AI-regulated world that humans are unable to control and direct may not be just a remote dystopia.

Agency is a basic human good and closely related to the conception of dignity [9] [25] [31] [32]. Acting in interaction with an AI-modified environment may change the basic conditions of human action and, therefore, the intrinsic value of human dignity. How can we conceptualize the changes in human agency and protect the terms of human dignity? The unpredictable nature of cooperation in the AI environment involves the danger that the categories of instrumental and technical goodness [31] will overrule intrinsic moral values.

Previous moral philosophical [9] and psychological research [19] [14] shows that three features characterize human agency: ableness, intentionality, and rationality. In the following section, we use these three concepts to analyze human–AI cooperation in order to detect and explore the possible impacts of AI systems on human agency and their effects on conceptualizations of human dignity.

3 Agency and Moral Agency

3.1 Conditions of Agency

For decades, academics have studied how human–computer interaction situations—involving, particularly, the sense of control humans have over actions and their consequences—may change the way in which we conceptualize human agency [14] [19] [2]. Although the discussion has been in many ways insightful, the results of these studies cannot be straightforwardly applied to human interaction with AI systems, and even less so to AI with designed learning capacities.

Traditionally, human–computer interaction has been understood as a form of bilateral communication. The problem with this approach is that it disregards the features in the relationship that give rise to conceptualizing it as a cooperative situation between a single actor or a group of human agents and an AI system. From the perspective of users, AI systems appear to be environments in which they cannot foresee what will take place or how to control the environment. The unpredictable nature of moves and countermoves is accentuated in systems where human and machine partners form a cooperative collective [18]. As AI systems and their functioning become more complex with interactive elements and learning capacities, the number of problematic issues related to human–computer interaction grows.

Human beings tend to personify their environment; boats are given names, and engines are cursed at for not working properly. It is then no wonder that we attach personified characteristics to complex AI systems. [22] [1] Does our proneness to anthropomorphize the world in which we act mislead us, to give artifacts and technology roles and positions that will eventually make us less human? To better grasp the issue at hand, let us first discuss the conditions for calling a performing capacity “agency.”

Aviation safety has improved immensely in virtue of various detection technologies and autonomous steering systems [4]. As these findings are translated into applications in the development of mass-manufactured motor vehicles, the role of human drivers is gradually decreasing. A fleet of self-driving cars in everyday use is no longer science fiction but an integral part of the prospect of improved road safety [15]. For a human being, such vehicles seem to make decisions and act upon them; as such, the question arises of whether we should view a self-driving car as an agent.

Historically, the concept of agency has been tightly tied to human beings and their capacities. There is, however, no consensus about the necessary or sufficient features defining the concept. It is safe to say that, in humans, agency is not a binary property but that its emergence takes place gradually during the physical, social, and psychological development of a child.

Even newborns, whose needs are provided for by parents and other caregivers, start to notice that what they do has an impact on those around them; for example, crying will usually summon someone to come to check on them, while gurgling or smiling will make the caregiver coo and caress them. Little by little, babies learn how to control their movements and the sounds they utter in a way that will enable them to make certain things happen. [24]

The same aspects that are crucial in child development reflect two features that are central to the conception of agency. An agent must be able to do things that will, in some way, change how things are in the world. Part of these functions relates to oneself (e.g., babies learn how to put their thumb in their mouth), but more importantly, these functions can influence the outside world. Being able to affect reality is a necessary condition for agency [9]. Let us call this the "ableness condition" (for examples of the use of this term, see Morriss 2002, 80–86) [20].

The second feature related to the notion of agency that is already visible in early childhood is the purposefulness of activities [24]. Our needs must be satisfied; one's wants and wishes indicate that there is a will to change how things are in reality. Even a rudimentary sense of oneself and a vague understanding of one's abilities form a basis for intentionality. An act is based on a goal that an actor pursues, using their activity as a means by which to reach it. Achieving almost any goal necessitates performing a series of acts that, together, form the action for pursuing or realizing an end. For an activity to be a manifestation of agency, the "intentionality condition" states that the activity must be a means designed for achieving the desired end.[9] [18]

Intentionality connects the aim and the action for achieving it, but this connection cannot be arbitrary. If there is no real possibility of reaching the goal or making its realization more probable, the activity is useless for furthering one's chosen end. To tie the goal to the action, agents must therefore be rational, in the minimal sense that they can discern fanciful ideas from realizable plans and have an idea of at least some of the actions that they are able to carry out. Let us call this requirement the "rationality condition." [9] [18]

Intentionality connects the aim and the action for achieving it, but this connection cannot be arbitrary. If there is no real possibility of reaching the goal or making its realization more probable, the activity is useless for furthering one's chosen end. To tie the goal to the action, agents must therefore be rational, in the minimal sense that they

can discern fanciful ideas from realizable plans and have an idea of at least some of the actions that they are able to carry out. Let us call this requirement the "rationality condition." [9] [24] [29

3.2 Features of Moral Agency

After having sketched central conditions of agency, let us consider the features of moral agency. As a starting point, we can use Thomas Aquinas's first principle of practical reason, a guiding norm of action (i.e., morality): "Bonum est faciendum et prosequendum et malum vitandum," meaning, "Good must be done and pursued and evil avoided" (Summa Theologiae I-II, 94:2) [27].

The main features central to moral agency serve as the presuppositions for the first principle of practical reason, as it is not possible to follow the norm without them. First, the principle expresses a positive and a negative norm. The positive norm exhorts the agent to do certain types of deeds and pursue certain types of ends; the negative norm tells the agent to avoid certain types of deeds and, implicitly, certain types of ends. The negative norm could also be expressed in a more categorical way by forbidding the agent to carrying out certain types of deeds and pursuing certain types of ends.

Second, morality presupposes that agents pay attention to discerning the types of actions they are supposed to do—that is, discerning good actions from those they must avoid doing, and good ends worth pursuing from those that are evil or bad. Third, agents must commit themselves to following both the positive and the negative norm in their acting as they choose the goals and means of their actions. It is the willingness to do so and to commit oneself to following the first principle of practical reason that makes an agent moral.

What Thomas Aquinas's definition does not express is the social nature of morality. An implicit presupposition in morality is that humans naturally aim at their (subjectively supposed) good and do not need any specific norm to motivate them to do so. What makes ends and deeds morally good or evil is their effect on other people and their wellbeing [25]. An important part of the social aspect of morality is that moral agents can be blamed and praised for what they do and what their actions cause. The social practices of blaming and praising indicate that the agents are being held responsible for what they do and cause by their deeds.

The central features of moral agency tie actors together into a community of other moral actors in which they are bound to each other, forming an intertwining web of relationships in virtue of the effects of their actions on each other and everyone [9]. Communities form and dissolve for various reasons; in between, they exist for some purpose that ties the members to each other and to the community [29]. Acting based on this purpose and being mutually connected create communication and meaning—a group culture that takes place in time, forming a common history for the members. In this way, belonging to and being part of any community is historically constituted. [8]

"Moral responsibility" is acknowledging one's role in the moral community and committing oneself to doing one's share as a part of the whole. Being morally responsible involves moral agents identifying themselves as givers and receivers of various goods within a network of mutual interdependency. To adopt the role and viewpoint of

a moral agent implies that one is willing to widen one's individual perspective and consider things from a universalizable moral point of view.

To sum up the central features of moral agency, we can say that moral agents (are willing and challenge themselves to) discern what is good from what is evil, and they commit themselves to doing and pursuing good and avoiding evil. In doing so, they commit themselves to extending their consideration not just to themselves but also to others. They identify themselves as members of a socially and historically constituted moral community within which they take and bear responsibility for their actions and the consequences of their actions. In which sense, then, can we speak of AI as having features of moral agency? Can human–AI interaction change the features of moral agency in humans?

3.3 Do AI-systems Count as Moral Agents?

The development of AI systems has made the boundaries between instrumental technologies and human agents opaque. Human agents tend to interact more efficiently with AI systems when they perceive them to be humanlike entities. Social robots and bots developed for social contexts resemble human actors both in their functions and in their given identities: designers give names to technologies operated by social AI and share narratives of their history. Learning machines develop their own functioning. This development has led to claims that we should attribute agency to complex self-learning AI systems and assign at least partial responsibility to AI agents functioning in interaction with human agents. [8] [5] [1]

Hakli and Mäkelä (2019) present a multifaceted argument against the view that AI systems are moral agents that can be held responsible for their actions. They introduce various arguments in the current discussion on the subject but maintain that not even future developments of the technology could furnish AI systems with the conditions necessary for moral agency. Their ground for arguing against extending moral responsibility—a feature central to moral agency—to AI systems and robots is that they lack—and will always lack, no matter how intricate their technology—autonomy and reflective self-control. [8]

Such a view is based on Alfred R. Mele's (1995) [17] contention that both autonomy and responsibility are features that gradually develop during the lived history of an agent. It is not just the intentional attitudes, values (i.e., the view of certain behaviors as intrinsically good, or "pro-attitudes") and capacities of an agent that matter but the causal history through which they were formed. In contrast, every AI system and robot, those with in-built learning capacities included, are initially brought into existence—or produced—by someone who programmed them. It then follows that the values and preferences displayed as the pro-attitudes that direct or determine the goals of an AI system or a robot can be traced back to the initial engineering design of the system. From a moral point of view, engineering counts as manipulation that undermines moral autonomy. Even when AI systems display pro-attitudes that direct their functioning and, in this sense, simulate goal-oriented agents, their intentionality is not authentic for the reason of its origin as a programmed propensity. As such, AI systems cannot be

held responsible for the outcomes of their functions and therefore cannot count as moral agents either. [8]

What Hakli and Mäkelä's conclusion implies in terms of moral agency and human responsibility is that the complexity and self-learning features of machines and artifacts do have an impact on responsibility assessment, opening such considerations to responsibility gaps and making it harder to determine where both praise and blame lie. The way forward, then, is not to attribute moral responsibility to AI systems but to assess anew the nature and boundaries of human moral agency.

3.4 Is Human-AI Interaction Collective Action?

Another approach to considering whether AI systems can be held morally accountable is to examine human–AI interaction. An important part of human agency takes place in the context of multiple agents. A strong theme in an ongoing discussion concerns whether human interaction with AI systems counts as collective action involving shared or joint intentional states or cooperation. [18] [266]

Whatever stance we take to the matter, AI systems and their various applications form a growing number of complex webs of human–machine systems: the AI systems continuously collect information about their users and apply it as material for their learning algorithms that again change their functioning in relation to the users, and so on.

It is hoped that AI-assisted data analysis will not only help in achieving goals set by human users of AI systems but that the systems themselves will pave the way for new findings, such as in medical diagnostics and the prevention of social issues [3]. Different types of data records could be used to improve both health and social wellbeing on a national, and even global, level. Not all researchers are equally enthusiastic about such prospects. Some claim that this new type of technology (i.e., AI systems) necessarily changes our view of humanity and morality and, therefore, also changes our conceptual terms of human action [23] 30.

Hakli and Mäkelä [8] consider whether human–machine interaction could be viewed as joint or collective action, as they discuss the possibility of holding AI agents at least partially responsible for the outcomes and consequences of their workings. According to their argument, the causal history of an AI system makes it implausible to regard the system as a moral agent capable of moral responsibility. They refute Latours's (2005) [12] view that collective agents, some of which are human beings and some artifacts, could be ascribed partial responsibility.

Hakli and Mäkelä back up their claim by referring to the philosophy of social action. Here, the definition of "joint action" reads as follows: two or more individuals perform a joint action if each of them intentionally performs an individual action but does so with the (true) belief that in performing the action, they will jointly realize an end that each of them has. The definition ascribes joint responsibility to the individuals partaking in the action. The notion of joint action implies that every group member is individually morally responsible for the joint action and its outcomes but also individually responsible, jointly and interdependently with the other members of the group. [8] [29]

As AI systems cannot be held morally responsible agents, they also do not fulfill the conditions of joint responsibility.

Could programming AI systems to perform morally favorable preferences serve as a counter claim to the negative view concerning responsibility and moral agency in AI systems? Let us consider a simple example. There are AI-monitored systems for assigning doctor appointments to patients. Such a system's designers can program the system to discern the patients whose need for medical examination or care is urgent from those whose condition allows for a longer waiting time before the appointment. It is possible to include different types of parameters in the preference algorithm, such as the patients' age, social status, and indicators concerning vulnerability, to increase the moral sensitivity of the algorithmic principles concerning the administration of consultation times.

Adding such features into the algorithm would improve the monitoring of appointments in medical care in terms of fairness—which is a moral characteristic—but it would not make the AI system morally responsible for placing the patients in a preferential order. Those who feel they have unjustly been left waiting in the patient queue would, rightly, blame those who designed the code for the monitoring system rather than the AI application itself.

From the viewpoint of an AI system, the programming that produces the morally favorable features is in no way different from any other piece of code in its algorithm. It would be justified to call such an appointment monitoring application a well-functioning AI system, but it would not make the system a moral agent.

So far, there are no convincing arguments for assigning moral agency or responsibility to AI systems or robots, and if Hakli and Mäkelä are correct, there are conceptual reasons that prevent us from doing so notwithstanding any future developments in AI. For this reason, human–AI interaction is not genuinely or fully collective action.

4 Moral Agency within Human-AI System Interaction

Let us now scrutinize how acting within an AI system or AI-monitored scheme may affect the notion of human agency by relying on the three characteristics of agency featured in section 2 (i.e., the ableness condition, the intentionality condition, and the rationality condition).

Technology is always designed to serve some purpose; it is therefore never value neutral. The ends that artifacts are supposed to fulfill or further become visible through their uses and hoped-for effects. Technology plays a causal, not an intrinsic, role even in interactive settings between humans and AI systems. [8]

As part of technology, AI is designed to assist human beings in their various enterprises. Machines are there to do the work that is too hard—in any sense of the word—for people. During the era of industrialization, motorized engines have replaced human labor in a growing number of tasks. Machines take care of chores that involve the use of physical power or repetitive actions. When machines take over human tasks, at each phase of the process, there are people who lose their jobs and positions. In the beginning of the process of industrialization, the development of technology replaced those doing

manual labor. However, during recent decades, AI systems have begun to replace the skilled human workforce; in the coming years, an increasing number of professions are likely to become futile.

So far, machines have not become superior to human labor in terms of its most valued human features, and it has been consoling to think that human beings still take care of the highest functions of any activity and that machines only perform auxiliary tasks. The development of autonomous and self-leaning AI systems changes this. For example, the improvement of aviation safety now strongly relies on lessening the impact of human perception and decision making in managing an aircraft [11]. The same development is already in progress for sea and land traffic [13] [16]. The latest findings show that AI-based screening programs are able to detect tumors that doctors do not notice [3].

There is no prospect any time soon of a limit to the things that AI systems can do better than human agents. So, at what point does widening the range of technology become a form of paternalism that inhibits people from performing tasks for the sake of protecting their own interest? It is likely that being second to a machine in an increasing number of fields of expertise will affect how human beings regard their own abilities, and the use of self-learning AI will accelerate this process. The more human beings—who get ill and grow old, err and fail—can be replaced by robots and AI systems, the more they will look like a source of unnecessary cost.

Even artists may notice that robots can take over their jobs. Musical robots can be programmed to have a touch for any genre of music; from a large set of data about popular hits, they can produce an endless variety of new ones. [22] There is no characteristic as such that could be called “creativity” that would help to discern a human-composed piece from one produced by AI [28]. Linguistic programs are already versatile enough to complete many tasks, and the speed with which translation machines have improved predicts that, someday, AI systems will master linguistic skills that have traditionally been thought to make people unique. There are fewer and fewer human abilities that make human beings superior to machines. What, then, is the locus of human agency?

AI-regulated environments also modify the conditions of human intentionality. Unknowingly to the users, systems run by algorithms fix the available goals or set them in a preferential order. For example, recommender algorithms used by social media platforms, music and movie streaming companies, online marketing agencies, and dating applications, collect user data such as previous online behavior and preferences and make recommendations or dictate what content users see [16]. Consequently, the users may cherish false expectations, pursue unachievable goals, or disinvest their time and energy, just because they are not fully aware of the algorithmic terms of the system.

Such developments are often unintentional; it is not a part of the concept of AI to curb human agency but to enhance it. The effects of activities in AI-regulated environments are unpredictable, but they may have wide-ranging implications in terms of how we perceive ourselves and treat other humans as intentional agents. There is already psychological evidence that performing computer-assisted tasks weakens the experience of being fully in charge of one’s actions [2]. People show different personality traits when communicating with an AI than when interacting with another human [21].

Cooperating with AI and robots may therefore deteriorate the conditions that accentuate the feeling of moral responsibility in human agents.

Reliable, relevant, and wide-ranging information is important for any meaningful and responsible—and, in this sense, rational—decision-making process. Multinational giants such as Google, Facebook, and Twitter apply algorithms that have a huge impact on national and international politics and economies, as well as on ordinary people's private lives [16]. The ways in which the systems direct and manage the flows of information may remain a mystery even to their designers. There are already ample examples of how algorithms can be prejudiced in terms of the sources of available knowledge and how they can boost disinformation, thereby distorting the rational decision-making processes of human agents.

As the above examples show, it is not difficult to find cases of AI-regulated environments in which even the present use of AI technology impairs the conditions of human agency. It is, therefore, not exaggerating the risks to fear that such developments may weaken the terms on which our sense of human dignity and the institutions that support it rely.

5 Conclusions

There is already evidence of the effect that cooperation with AI systems has on humans' perception of the conditions of their agency. Great expectations concerning the instrumental benefits of AI-enhanced activities and their economic value direct the development of the AI industry. This may overshadow the implicit effects that living and working in various AI environments have on human agency and moral responsibility.

Listing ethical principles for AI development and use is not enough; they are abstract ideals, hard to apply in actual design and use. It is necessary, therefore, to obtain a deeper understanding of how technological progress takes place and how it is intertwined with economic and political realities in different types of societies and in the global community.

No matter how skillfully, diligently, and creatively AI systems can be designed to work, moral responsibility remains a human characteristic. It is therefore our duty to develop AI to serve morally good ends and purposes. Protecting and strengthening the conditions of human agency in any AI environment is part of this task.

References

1. Brożek B. & Janik B.: Can Artificial intelligence be moral agents? New Ideas in Psychology, 54, 101–106 (2019).
2. Ciardo F., De Tommaso D., Beyer F., Wykowska A.: Reduced Sense of Agency in Human-Robot Interaction. In: Ge S. et al. (eds.) Social Robotics. ICSR 2018. Lecture Notes in Computer Science, vol. 11357, pp. 441–450. Springer, Cham (2018).
3. Esteva, A., Kuprel, B., Novoa, R. et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542**,** 115–118 (2017).
4. FAA: Operational Use of Flight Path Management Systems. Final Report. FAA, (2013). https://www.faa.gov/aircraft/air_cert/design_approvals/human_factors/media/oufpms_report.pdf, last accessed 2020/01/28.
5. Fossa F.: Artificial moral agents: moral mentors or sensible tools? Ethics and Information Technology 20, 115–126 (2018).
6. Grace, K. & Salvatier, J. & Dafoe, A. & Zhang, B. & Evans, O.: Viewpoint: When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research 62: 729–754, (2018).
7. Gunkel, D. J.: The Machine Question: Critical Perspectives on AI, Robots, and Ethics. MIT Press. Cambridge MA (2012).
8. Hakli, R. & Mäkelä, P.: Moral responsibility of robots and hybrid agents. The Monist 102(2), 259–275, (2019).
9. Hallamaa, J.: Yhdessä toimimisen etiikka [The Ethics of cooperation]. Gaudeamus, Helsinki (2017).
10. Kant, I.: Grundlegung zur Metaphysik der Sitten. Schriften zur Ethik und Religionsphilosophie. Erster Teil. Wissenschaftliche Buchgesellschaft, Darmstadt, (1983).
11. Landry, S. J. & Karwowski, W.: Advances in Human Factors and Ergonomics Series: Advances in Human Aspects of Aviation. CRC Press LLC (2012).
12. Latour, B.: Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, New York (2005).
13. Leikas, J. & Koivisto, R. & Gotcheva, N.: Ethical Framework for Designing Autonomous Intelligent Systems. Journal of Open Innovation: Technology, Market, and Complexity 5(1): 18–30, (2019).
14. Limerick H., Coyle D. and Moore J. W.: The experience of agency in human-computer interactions: a review. Frontiers in Human Neuroscience 8:643, (2014).
15. Lin, P., Abney, K. & Jenkins, R. (eds.): Robot ethics 2.0: From autonomous cars to artificial intelligence. Oxford scholarship online (2017).
16. Marr, B.: Artificial intelligence in practice: How 50 successful companies used AI and machine learning to solve problems. Wiley, Chichester, West Sussex, UK (2019).
17. Mele, A. R.: Autonomous Agents: From Self-Control to Autonomy. Oxford University Press, New York (1995)
18. Misselhorn, C.: Collective agency and cooperation in natural and artificial systems. In: Misselhorn, C. (ed.) Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation, pp. 3–24. Springer International Publishing, Switzerland (2015).
19. Moore, J. W.: What is a sense of agency and why does it matter? Frontiers in Psychology 7:1272 (2016).
20. Morriss, P.: Power. A philosophical approach. 2nd edition. Manchester University Press, Manchester (2002).

21. Mou, Y. & Xu, K.: The Media inequality: Comparing the initial human-human and human-AI social interactions. Computers in Human Behavior, 72: 432-440 (2017).
22. Needham, J.: We are the robots: is the future of music artificial? FACT Magazine, https://www.factmag.com/2017/02/19/we-are-the-robots-could-the-future-of-music-be-artificial, last accessed 20/01/28.
23. Ollila, M.-R: Tekoälyn etiikkaa [Ethics of artificial intelligence]. Otava, Helsinki (2019).
24. Rochat, P.: Others in mind: Social origins of self-consciousness. Cambridge University Press, New York (2009).
25. Smith, C.: To Flourish or destruct: A personalist theory of human goods, motivations, failure and evil. The University of Chicago Press, Chicago (2015).
26. Strasser, A.: Can artificial systems be part of a collective action? In: Misselhorn, C. (ed.) Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation, pp. 205–219. Springer International Publishing, Switzerland (2015).
27. Thomas of Aquinas: Summa Theologiae. Pars prima et prima secundae. Marietti, Torino (1952).
28. Thompson, Clive: What will happen when machines write songs just as well as your favorite musician? Mother Jones, https://www.motherjones.com/media/2019/03/what-will-happen-when-machines-write-songs-just-as-well-as-your-favorite-musician, last accessed 2020/01/28.
29. Tuomela, R.: The Philosophy of sociality: a shared point of view. Oxford University Press, Oxford, England (2007).
30. Visala, A.: Tekoälyn teologiasta [Remarks on theology of artificial intelligence]. Teologinen Aikakauskirja, 123(5), 402–417 (2018).
31. von Wright, G. H.: The Varieties of goodness. Routledge and Kegan Paul, London, England (1968).
32. Zdravkova, K.: Reconsidering human dignity in the new era. New Ideas in Psychology, 54, 112–117 (2019).